


RESEARCH

Open Access



Fine-grained precise-bone age assessment by integrating prior knowledge and recursive feature pyramid network

Yang Jia^{1,2,3*} , Xinmeng Zhang^{1,2,3}, Hanrong Du^{1,2,3}, Weiguang Chen^{1,2,3}, Xiaohui Jin⁴, Wei Qi⁵, Bin Yang⁴, Qiujuan Zhang⁵ and Zhi Wei⁶

*Correspondence:
jiayang@xupt.edu.cn

¹ School of Computer Science & Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, Shaanxi, China

² Shaanxi Key Laboratory of Network Data Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an 710121, Shaanxi, China

³ Xi'an Key Laboratory of Big Data and Intelligent Computing, Xi'an 710121, Shaanxi, China

⁴ Department of Radiology, Xi'an Honghui Hospital, Xi'an 710054, Shaanxi, China

⁵ Department of Radiology, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710004, Shaanxi, China

⁶ School of Computer, New Jersey Institute of Technology, Newark, NJ, USA

Abstract

Bone age assessment (BAA) evaluates individual skeletal maturity by comparing the characteristics of skeletal development to the standard in a specific population. The X-ray image examination for bone age is tedious and subjective, and it requires high professional skills. Therefore, AI techniques are desired to innovate and improve BAA methods. Most of the BAA method use the whole X-ray image in an end-to-end model directly. Such whole-image-based approaches fail to characterize local changes and provide limited aid for diagnosis and understanding disease progress. To address these issues, we collected and curated a dataset of 2129 cases for the study of BAA with fine-grained skeletal maturity level labels of the 13 ROIs in hand bone based on the expert knowledge from TW method. We designed a four-stage automatic BAA model based on recursive feature pyramid network. Firstly, the palm region was segmented using U-Net, followed by the extraction of multi-target ROIs of hand bone using a recursive feature pyramid network. Given the extracted ROIs, we employed a transfer learning model with attention mechanism to predict the skeletal maturity level of each ROI. Finally, the bone age is assessed based on the percentile curve of bone maturity. The proposed BAA model can automate the BAA. In addition, it provides the detection result of the 13 ROIs and their ROI-level skeletal maturity. The MAE can reach 0.61 years on the dataset with the labeling precision of one year. All the data and annotations used in this paper are released publicly.

Keywords: Bone age assessment, Multi-target detection, Prior knowledge, Recursive feature pyramid network, Transfer learning

1 Introduction

To reflect the maturity of the body, bone age is more accurate than biological age. Bone age assessment (BAA) is a technology which evaluates the maturity of individual bones by measuring the differences between bone age and biological age according to the common characteristics of bone development of a specific population. It is the most accurate and objective method for clinical evaluation of individual development. It plays an important role in the diagnosis of pediatric endocrine problems and children's growth disorders [1]. In addition, it is also a significant index for identifying the real age of

suspects in juvenile delinquency cases, the age of athletes in sports competitions [2], height prediction, and selection of athletes. Usually, an X-ray image of a hand is taken during the assessment. As time goes by, the shape, color and the appearance of the bones will change, and the doctor will check the whole image or each key bone of the hand to assess the bone age.

Traditional BAA process is tedious, while it is subjective and exhibits substantial discrepancy [3], leaving room for innovated improvement. Since the BAA standards [4–6] have distinct and standardized descriptions, BAA is very suitable to be fully automated by machine learning methods. In the early stage of computer-aided diagnosis (CAD), an expert system was mainly used to estimate bone age with image processing and pattern recognition [7, 8]. In recent years, with the development of deep learning technology, numerous end-to-end BAA algorithms and models based on the deep neural network have emerged. In these end-to-end models, users input an X-ray image into the model and get the bone age as the output directly [9–18]. Imagining that, once you take an X-ray image, you will get a bone age within several seconds, which will save a lot of time and labor. In these studies, for each sample image in the dataset, there is an age assessed by a professional radiologist and attached as the label. However, clinicians and radiologists think that the scoring method should provide more fine-grained and precise results, and they want to see the development of each bone in hand, rather than just get one single bone age number. If the software can automatically mark the bone development changes of a patient in time sequence, it is much more valuable to analyze the patient's progression and adjust the treatment plan [15]. In order to get the fine-grained and precise bone maturity evaluation result, all of the regions of interest (ROIs) in hand based on the medical standards, such as TW method and RUS-CHN [5, 6], need to be detected and analyzed one by one and this precise analysis is more challenging because of the multiple small bone detection task and the lack of precisely annotated ROIs.

In fact, there exists a multi-bone detection framework for BAA [19–22]. Faster R-CNN are commonly used to detect the ROIs, and then the features of each ROI are used as the input of a regression network which consists of a variable set of fully connected layers followed by a one-neuron output layer providing a bone age estimation. Nonetheless, there are some defects of the studies. First, the statistical evaluations of the ROI detection result are not rigorous. Second, not all the 13 ROIs in the BAA scheme based on the TW3 method are incorporated. Another major limitation of these works is that the maturity of each ROI was not provided, and the prior medical knowledge of BAA was not fully considered in the studies.

In this study, to address the aforementioned problems, we work together with two hospitals and build a dataset of 2129 X-ray images. 1015×13 ROIs were annotated for this study. Moreover, we labeled 100 images for palm segmentation. Based on the Chinese version of the TW method [5], we proposed a fine-grained and precise BAA method called precise-BAA to process the X-ray image of hand bone based on prior knowledge and a recursive feature pyramid network. To model the 13 ROIs of the hand bone, we designed a four-stage detection and regression BAA framework to handle the fine-grained bone analysis. The framework is shown in Fig. 1. The palm is segmented from the original X-ray image to exclude the interference and benefit the deep learning model in the next stage. Different from other solutions, our method detects 13 small ROIs with

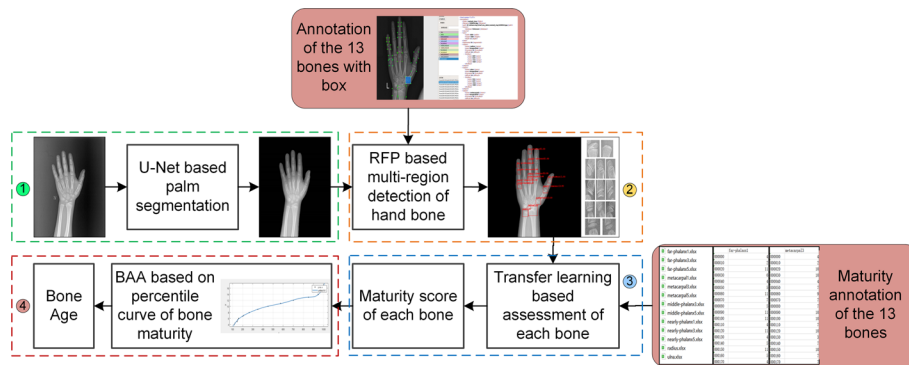


Fig. 1 Illustration of the proposed method: (1) the U-Net module completes image preprocessing and palm segmentation; (2) fine-grained small ROI detection based on recursive feature pyramid. (3) Maturity of each bone assessment based on transfer learning. (4) Bone age calculation based on the percentile curve of bone maturity

a recursive feature pyramid network in the second stage, and evaluates each bone based on the knowledge from TW method. In the third stage, a regression model transfer from natural image recognition is built to assess the level of each ROI, and the attention scores from other bones are used to assist the local evaluation of the ROI extracted at the fine-grained stage. Finally, we calculate the bone age based on the percentile curve of bone maturity.

The contributions of the proposed methods can be summarized as follows:

1. A multi-bone detection network with recursive feature pyramid network and cascade R-CNN is proposed to detect the 13 ROIs for BAA. The experimental result shows impressive performance and our method outperforms state-of-the-art sub-region detection techniques in BAA.
2. A four-stage BAA framework is proposed and the four parts are independent and concurrent. A fine-grained dataset with 100 labeled images for palm segmentation and 1015×13 bone maturity level annotations is provided, which is different from the dataset used for training in the state-of-the-art research. It serves as the bridge for expert knowledge and deep learning model.

The paper is organized as follows. Section 2 discusses the state-of-the-art related research. In Sect. 3, we present the preprocessing of the X-ray image of hand bone. Section 4 introduces fine-grained multi-bone detection based on recursive feature pyramid network and cascade R-CNN. Section 5 presents the transferred model with self-attention for ROI maturity level prediction and final BAA. A comprehensive evaluation is provided in Sect. 6.

2 Related works

2.1 Basic medical methods of BAA in this study

There are two categories of medical methods for BAA: G-P method [4] and the TW scoring method [5]. The G-P method describes the standard atlas of different age stages, and lists the maturity indicators of the bones in detail. Bone age is given by

comparing the X-ray image to the standard atlas. Nowadays, many hospitals are still using this method. However, it needs to consider various maturity indicators comprehensively in the whole comparison, and it is very subjective. Another one is called TW scoring method [5]. Many other bone age scoring methods, such as CHN [5] and RUS-CHN [23] are both derived from TW scoring method [5]. TW method is widely accepted all over the world and many countries have established their own bone age standards based on TW, which makes BAA more accurate and effective. The basic idea is to select several bones in the hand and wrist for evaluation, as shown in Fig. 2. The 13 bones used in this study are marked with yellow numbers in Fig. 2a, and the name list of the bones can be found in that figure. Figure 2c shows the enlarged ROIs. The ROI is fine-grained and the estimation should be precise enough to handle these local regions. According to the rules of estimating skeletal indicators, the whole development process of each bone is divided into several stages and a score was given to each bone stage, and the radiologists will grade all the 13 bones one by one. Then a total score of the skeletal maturity of all the bones is calculated. Finally, the bone age is determined by checking the percentile curve of bone maturity.

The X-ray image of a hand can be divided into three regions: background, soft-tissue area and bone. In the early stage of CAD BAA research, researchers tried different kinds of segmentation methods to remove the soft tissue area of the hand, and just keep the bone area for BAA [24]. Due to the gray overlap of some pixels between the bone region, soft tissue area, and background, especially in some epiphyseal regions, as shown with the red arrow in Fig. 2b, the gray-level distributions of different parts of the hand bone are different, and furthermore, the gray levels of different tissues overlap. It means that the threshold method is not suitable for hand bone segmentation. The clustering method ASM (active shape model) [25] and level set methods have also been studied, but the performance still needs further improvement.

With the development of deep learning techniques, quite a few deep learning-based methods have been proposed these years [26, 27]. They can be roughly divided into the following two categories based on their strategies.

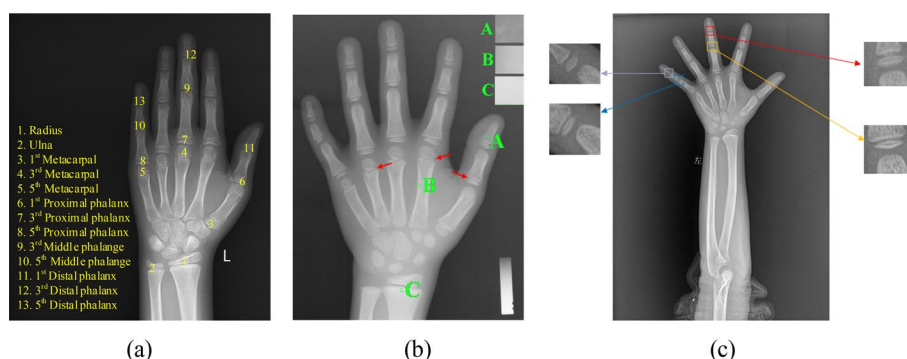


Fig. 2 Example of bones selected in the proposed method. **a** 13 ROIs for BAA. **b** Red arrow marks the soft tissue area with great difficulty in segmentation, and green sampling points mark the gray distribution of bone. **c** Enlarged ossification centers in the X-ray image

2.2 End-to-end deep learning-based BAA

The North American Radiology Society (RSNA) has hosted the famous Pediatric Bone Age Machine Learning Competition Challenge [28], which attracts a large number of researchers to participate in the competition and greatly promotes the development of automatic BAA technology. In recent years, our research group has also jointly carried out several research works with the Xi'an Honghui hospital and the Second Affiliated Hospital of Xi'an Jiaotong University, and published a series of research papers [11–15]. The general framework of automatic BAA nowadays is based on the deep learning method [3, 10, 28–30]. There are two common deep learning algorithm frameworks of BAA. One is to treat the hand as a whole. The other one follows the TW scoring theory, and processes the key regions separately. The first one is generally divided into three steps: input image preprocessing [9, 10, 12, 16–18] (not shown in the figure), feature extraction based on convolutional neural network, and bone age obtained by using regression network. The preprocessing part usually includes equalization of gray value of X-ray image, cropping, angle correction, and data augmentation [9, 31]. These auxiliary operations play an important role in improving the performance of the algorithm.

2.3 Subregion-based BAA

The second strategy is extracting the key regions of a hand using a region proposal network (RPN or Faster R-CNN) and evaluate each region separately [19–22, 32]. TW3 and CNN-based methods are integrated together while utilizing deep learning. Liang et al. [19] proposed a Faster R-CNN-based model to detect the ossification centers of epiphysis and carpal bones to evaluate bone age. For each patient, the bone age value annotated by the professional radiologists was also used as the label of all ossification centers, and they ignored the inconsistency of maturity levels of different ROIs. Bui et al. [20] proposed a BAA approach by integration of TW3 method and deep convolution networks based on extracted ROI-detection and classification using Faster R-CNN and Inception-v4 networks, respectively. Six ossification centers are annotated in their dataset, and finally, they used support vector regression to estimate the age. However, the 6 ROIs was simplified from the 13 ROIs in TW3 and detection of the 6 larger ROIs is much easier than detection of the original 13 ROIs. It lacks evidence that the 6 ROIs are enough for BAA. In [21], an automated system for BAA that mimics and accelerates the workflow of the radiologist was proposed. The system was based on Faster R-CNN, which detected 17 regions from the 6 ossification areas (DIP, PIP, MCP, radius, ulna, and wrist). The performance was measured by the average precision of the Intersection over Union (0.5IoU) for the 6 ossification areas [33] and a detected region is considered as a good match if the overlapped area accounts for at least 50% region. It is a lax measure of subregion detection for BAA. Then a gender and region-specific regression network was used to estimate the bone age. Because of lacking image preprocessing, for some low contrast and high noise X-ray images, there are lots of missing ROIs. There are 6 different models for each gender and the final bone age was given by 6 evaluations and the composite bone age for the subject was not provided. Compared to the end-to-end methods [9, 10, 12, 16–18], these

region-based methods utilize expert knowledge to improve the accuracy of BAA and provide more information about the bone age related regions and visual appearance. However, several challenging problems remain to be addressed:

1. The ROIs detection is not efficient enough for bone maturity assessment. The missing ROIs will affect the regression severely.
2. The annotation of the ROIs is not consistent with the medical standards and the fine-grained precise features of the bones are not considered.
3. The prior knowledge of bone development has not been infused into the BAA system effectively.

3 Methods

In this work, our goal is to recognize all the 13 ROIs in hand bone and evaluate each of them to auxiliary BAA, because the targets are similar in appearance, as shown in Fig. 2c. Among the 13 ROIs, the two epiphyseal areas in the yellow box and the red box, the areas in the blue box and purple box are very close in appearance. Moreover, the proportion of 13 ROIs in the image is also small, which makes this multi-target detection task more challenging. Therefore, we proposed this four-stage BAA method to process the 13 ROIs precisely. This section describes the presented framework for bone-age estimation using computer vision-based scheme, which is shown in Fig. 1.

3.1 Pre-processing of the X-ray image of hand bone

The original images in the dataset include the hand bone, marks ('L'), background, and edge. If the detection and classification of the ROIs are directly carried out on the original picture, the inconsistency will greatly affect the result [34]. So, the images are pre-processed firstly to reduce the interference of redundant noise on the results.

3.1.1 Data preparation

The dataset used in this study is mainly from the Second Affiliated Hospital of Xi'an Jiaotong University, which provides 2153 X-ray images of adolescents' hands together with gender and bone age assessed by professional doctors. The age of patients in the dataset is 0–18 years. Since we are studying the BAA of adolescents, the cases older than 18 are abandoned, and finally, 2129 cases are left. The distribution of images in different ages is unbalanced because the needs of adolescents of different ages are different. The need for BAA is not usually found in the kids younger than 3 years, and the samples of this age group are less in the hospital. The samples in this group are less than that of the 4- to 7-year-old, 8- to 11-year-old, and 12-year-old groups. The skeletons of the other three groups are in the rapid development stage. Parents are more concerned about the potential of body development and maturation. Their needs for BAA are larger and the number of samples in the hospital is also larger.

There are white borders on the edge of the hand bone X-ray images, and the brightness of the images is different. The background around the hand area is darker, while the color close to the edges is lighter. The positions of the letter 'L' on the image is also not the same, the inconsistency of the original images will affect the model training.

Therefore, we preprocessed the original images to unify the background and remove the redundant part.

3.1.2 U-Net-based hand segmentation

Thirteen hand bones in the palm are used to assess the bone age, and we segmented the palm including 13 bones from the original gray images and provided standard samples for hand bone detection and bone maturity assessment. We used U-Net [35] for palm segmentation in the preprocessing stage. The main structure of U-Net adopts a skip connection that the information captured in the initial layers is fused to the later layers, which means that the lower semantic information extracted from the input is also used for the deconvolution layer and realizes the precise pixel-level positioning. U-Net combines low-level and high-level information together. The low-resolution information provides the basis for object classification, while the high-resolution information provides the basis for precise segmentation and positioning. We tried threshold and deep learning-based segmentation for palm, and U-Net-based method showed the best result.

3.2 Fine-grained multi-bone detection of hand bone

There are a number of existing target detectors that have been shown to perform well by using a two-look and think mechanism. Here we proposed a recursive feature pyramid network (RFP) [36] based multi-target detection of hand bone X-ray image, and the framework of the network is shown in Fig. 3. We proposed a recursive feature pyramid network (RFP) [36] based multi-target detection of hand bone X-ray image, and we connected all the feedback of FPN to the backbone network with ASPP (atrous spatial pyramid pooling) [37], so that the features trained by the backbone network could better fit the detection task. RFP [36] is based on the feature pyramid network [38], which connects additional feedbacks to the backbone network. This recursive structure sequentially allows the resulting backbone to look at and think about the image multiple times. In addition, because RFP enhances the feature pyramid network recursively, it generates stronger feature representation ability (Fig. 3).

Moreover, switchable atrous convolution (SAC) is used to replace the standard convolution on the backbone network. Atrous convolution increases the receptive field of the network, and SAC can make the network more flexible [36]. SAC with different atrous rates captures the target of the different receptive fields, so the network learns a switch, which can adaptively adjust the convolution result of which receptive field to choose.

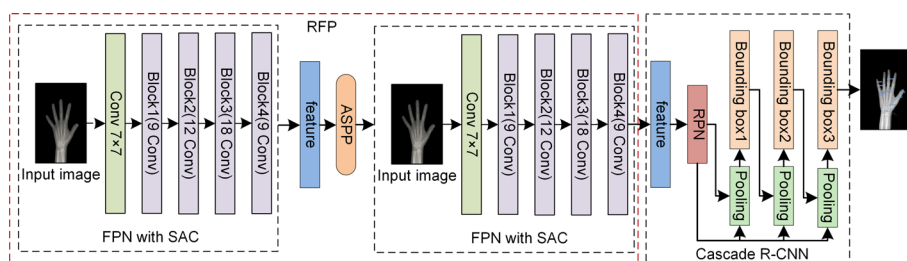


Fig. 3 Network of the proposed fine-grained multi-bone detection of hand bone

The switch function is more stable when global information is obtained. SAC is calculated as Eq. (1):

$$\text{Conv}(x, w, 1) \rightarrow S(x) \cdot \text{Conv}(x, w, 1) + (1 - S(x)) \cdot \text{Conv}(x, w + \Delta w, r), \quad (1)$$

where the input is x and the weight is w , r is the super parameter of SAC and atrous rate. Δw refers to the weight that can be trained, $w + \Delta w$ refers to the weight missed from the pre-training checkpoint. The switch function $S(\cdot)$ is composed of an average pooling layer of 5×5 and a convolution layer of 1×1 , which has a strong correlation to the input and location. After extracting the feature with RFP, the ROI bounding boxes are generated by RPN with cascade R-CNN, which is composed of a sequence of detectors trained with increasing IoU thresholds (0.5, 0.6, 0.7) to accomplish the high-quality detection [39]. RPN, SAC and cascade R-CNN greatly improve the performance of fine-grained bone detection.

3.3 Self-attention transfer network for BAA

After 13 ROIs were detected, we evaluated the levels of the ROIs based on TW method one by one. Two professional radiologists annotated 1015 images of hand bone X-ray images. For one image, all of the 13 ROIs are assigned levels indicating the maturity of bones. Details are in the supporting document. Transfer Learning [40] can infuse the prior knowledge accumulated in one task to another different but related problem. Transfer learning allows us to train new models with limited data, which will help us release the burden of annotations. Here knowledge learned from natural images is applied to hand bone X-ray images. In combination with available computing resources, we tested a variety of networks, including Inception-V3, ResNet, and Densenet, and designed a BAA model based on different network structures. Firstly, the feature extractor is designed, which used the Densenet, but the softmax layer at the top is removed. The dense connection structure in Densenet will increase the transfer between gradients, which plays the role of reuse feature, reduces the over-fitting due to the small sample size, and improves the accuracy of the model.

We found that the levels of the 13 ROIs in a hand X-ray image are correlated to each other, the Pearson correlation coefficients of the 13 ROIs are calculated and shown in Table 1. To further benefit from the correlation coefficients, we implement a simple and effective self-attention mechanism [41, 42] within the Densenet. As in Fig. 4, once we obtain our feature map, and are used to calculate the 13×13 attention map. The attention function on a set of queries is packed together into a matrix Q , the keys and values are also packed together into matrices K and V . We multiply the attention map and the feature map to get the self-attention feature map [43].

The structural schematic diagram of the hand bone target area grade evaluation model based on Densenet is shown in Fig. 5. After we got the predicted level of the 13 ROIs, the sum of the levels indicates the bone maturity of the case was calculated, as the X-axis in Fig. 6. Based on the 50th percentile curve of bone maturity from the RUS-CHN in Fig. 6, we can get the bone age of the case.

Table 1 Pearson correlation coefficients of the 13 ROIs

ROIs	1st distal phalanx	3rd distal phalanx	5th distal phalanx	1st metacarpal	3rd metacarpal	5th metacarpal	3rd middle phalange	5th middle phalange	1st proximal phalanx	3rd proximal phalanx	5th proximal phalanx	Radius	Ulna
1st distal phalanx	1	0.819	0.826	0.809	0.764	0.778	0.841	0.859	0.815	0.809	0.815	0.834	0.814
3rd distal phalanx	0.819	1	0.940	0.818	0.821	0.801	0.837	0.836	0.811	0.787	0.803	0.832	0.819
5th distal phalanx	0.826	0.940	1	0.833	0.840	0.831	0.843	0.848	0.827	0.791	0.805	0.850	0.832
1st metacarpal	0.809	0.818	0.833	1	0.787	0.809	0.875	0.846	0.864	0.869	0.866	0.832	0.874
3rd metacarpal	0.764	0.821	0.840	0.787	1	0.900	0.805	0.806	0.790	0.765	0.773	0.841	0.838
5th metacarpal	0.778	0.801	0.831	0.809	0.900	1	0.813	0.807	0.804	0.777	0.793	0.852	0.850
3rd middle phalange	0.841	0.837	0.843	0.875	0.805	0.813	1	0.887	0.870	0.848	0.879	0.864	0.872
5th middle phalange	0.859	0.836	0.848	0.846	0.806	0.807	0.887	1	0.850	0.838	0.849	0.869	0.844
1st proximal phalanx	0.815	0.811	0.827	0.864	0.790	0.804	0.870	0.850	1	0.852	0.876	0.844	0.853
3rd proximal phalanx	0.809	0.787	0.791	0.869	0.765	0.777	0.848	0.838	0.852	1	0.885	0.814	0.849
5th proximal phalanx	0.815	0.803	0.805	0.866	0.773	0.793	0.879	0.849	0.876	0.885	1	0.839	0.851
Radius	0.834	0.832	0.850	0.832	0.841	0.852	0.864	0.869	0.844	0.814	0.839	1	0.883
Ulna	0.814	0.819	0.832	0.874	0.838	0.850	0.872	0.844	0.853	0.849	0.851	0.883	1

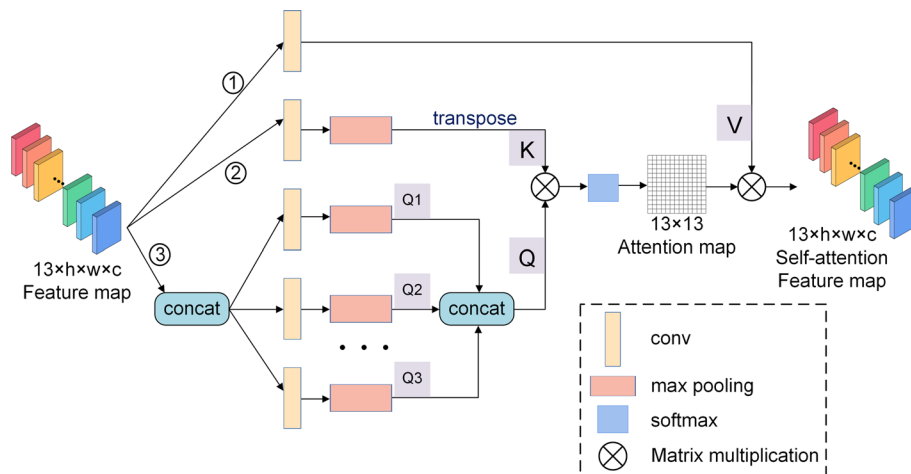


Fig. 4 Scaled dot-product self-attention calculation in this BAA model

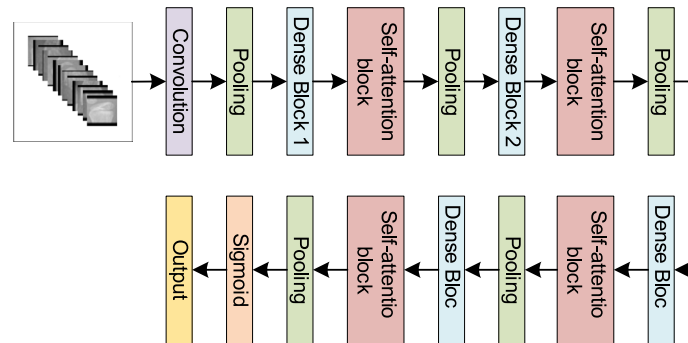


Fig. 5 The transferred model with self-attention for ROI maturity level prediction

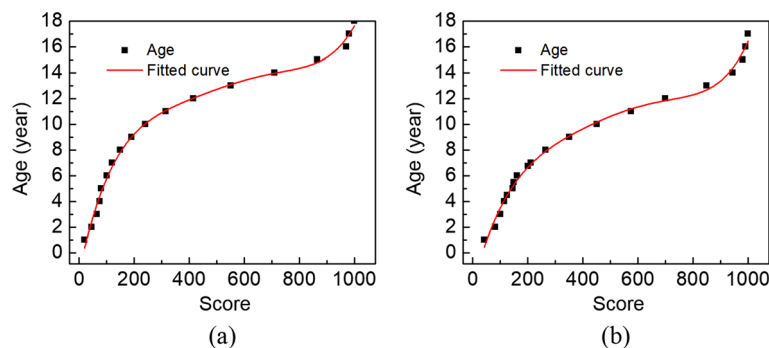


Fig. 6 The 50th percentile curve of bone maturity from RUS-CHN: **a** male, **b** female. The sum of the levels of the 13 ROIs is the “Score”, and with the curve we can get the bone age

4 Experiments and results

4.1 U-Net-based hand segmentation

100 samples are used in U-Net-based hand segmentation and the ratio of training and testing is 8:2. The parameters of the model are as these: the learning rate is set 0.01,

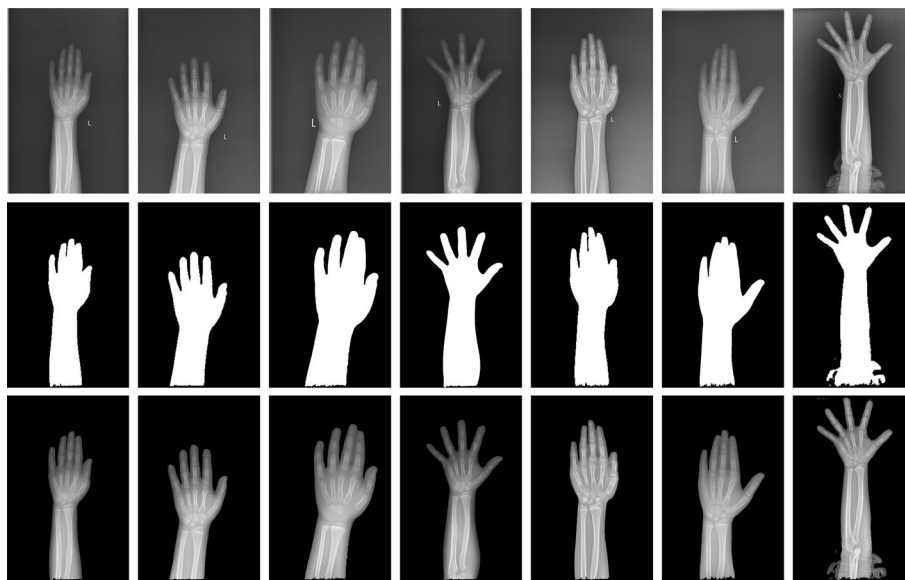


Fig. 7 Hand segmentation result based on U-Net model

0.001 and 0.005, the batch size is set 4, the momentum is 0.99, SGD is used as the optimizer, and the epoch is set 100. After 65 iterations, the loss curve of the training and the verification set tends to be flat, the loss tends to be stable around 0.025, the accuracy tends to be flat around 0.9801, and the model reaches the optimal state.

The segmentation results based on U-Net are shown in Fig. 7. Most of the hand masks show the palm region clearly, but some results are under segmentation. It is necessary to carry out post-processing of the predicted hand mask. By using the morphological method, extracting the maximum connection area and filling holes, a better hand binary mask can be obtained, as shown in the second row in Fig. 7. With this mask, we can get the final segmentation result, as shown in the third row in Fig. 7. Our experiments showed that removing the texts and inconsistent background would promote the detection performance greatly, which means the preprocessing is essential for the BAA algorithm.

4.2 Fine-grained multi-bone detection in hand bone

4.2.1 Evaluation settings

We use AP (average precision) and mAP (mean average precision) to evaluate the detection model. AP is for the evaluation for one ROI, and mAP is for the 13 ROIs. Equations (2)–(3) are for AP and mAP calculation:

$$AP_i = \frac{\sum \text{precision}_i}{N(\text{Total_Images})_i}, \tag{2}$$

$$\text{mAP} = \frac{1}{M} \sum_{i=1}^{i=M} AP_i, \tag{3}$$

where M is the classes of the objects. $N(\text{Total_Images})_i$ is the number of ROIs of class i .

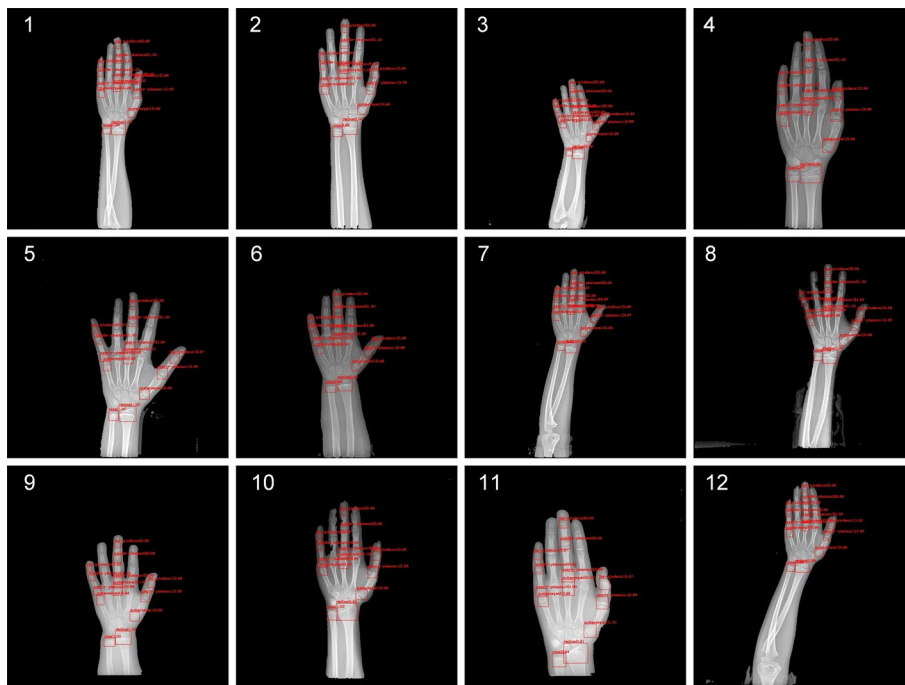


Fig. 8 Example outputs of ROIs detection in hand bone with RFP and SAC

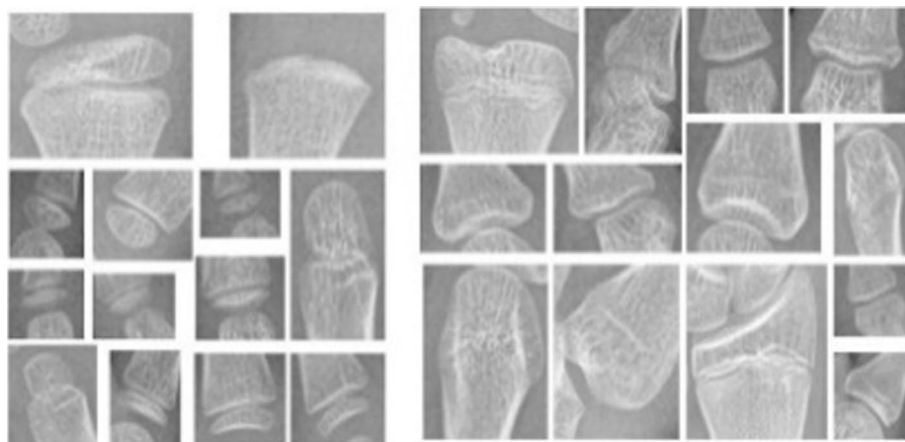


Fig. 9 Examples of the detected 13 ROIs in hand bone for bone maturity assessment

4.2.2 Multi-bone detection in hand bone with RFP and cascade R-CNN

We randomly selected 30% of the data from the hand bone dataset as the test set. We used PyTorch as the deep learning framework for the hand bone ROIs detection, which is implemented by Python3.5. The system is run in Ubuntu16.04 and accelerated by GPU (NVIDIA GTX 1080), and the memory is 8G. The batch size is set as 4, the number of iterations is set as 24, the SGD optimizer is used in the model, and the loss function is GIoU Loss [44]. What shocked us is the AP value of the 13 target regions finally obtained by the test set is all 1, and the final mAP is also 1. The test results are shown in Fig. 8.

No matter which age of the hand bone X-ray image is, with the proposed model, the 13 ROIs we need can be detected, and the positions are also very close to the ground truth. The model solved the fine-grained multi-target detection of ROIs, even the hardest part of the detection task, detection of the 3rd proximal phalanx, 5th proximal phalanx, 3rd distal phalanx, and 5th distal phalanx, as shown in Fig. 2c, they can be detected correctly. The extracted test results are shown in Fig. 9.

After we get the 13 ROIs of the hand bone, we can evaluate the maturity level of each ROI. We also built a faster R-CNN-based model [45] and a YOLO-based model to detect the 13 ROIs in the X-ray image. The comparison results are as follows.

4.2.3 Comparison of the multi-bone detection methods

Ren et al. [46] proposed a Fast R-CNN network using a region-generated network (RPN) to replace the selective search algorithm to generate candidate boxes. At the same time, RPN and the classification location prediction network share the convolution features, which greatly speed up the calculation, and with anchors of different scales adapting to different shapes of targets, the detection efficiency has been improved obviously. Considering the large scale of parameters, we used a pre-trained model with COCO and VOC datasets, and then 70% of our dataset are added to fine-tune the initial Faster R-CNN model. In the training stage, the initial learning rate is 0.001, the batch size is 16, the decay factor of weight is 0.0005. positive_overlap of RPN=0.7 (positive sample threshold), negative_overlap of RPN=0.3 (negative sample threshold), and NMS_THRESHOLD=0.7 (non-maximum suppression threshold). SGD optimizer is used. Three different iteration numbers are set, and the test results obtained after training are shown in Fig. 10.

The highest mAP is 0.87, however, the AP values of the 3rd proximal phalanx, 5th proximal phalanx, 3rd distal phalanx and 5th distal phalanx are lower than other ROIs. As shown in Fig. 2a, the 3rd proximal phalanx and the 5th proximal phalanx are labeled 7 and 8, these two regions are close to each other. Based on the TW method, the annotations are shown in Fig. 11. Green boxes are the 3rd and the 5th proximal phalanx, red

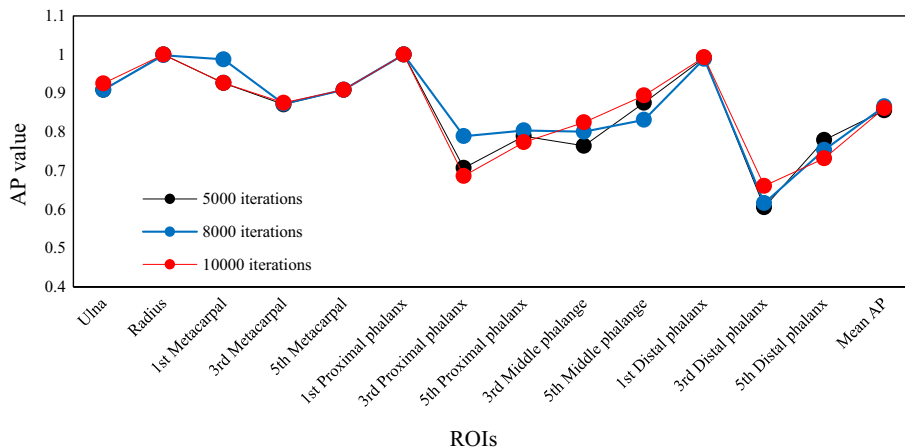


Fig. 10 AP values of the ROIs with the faster R-CNN-based model

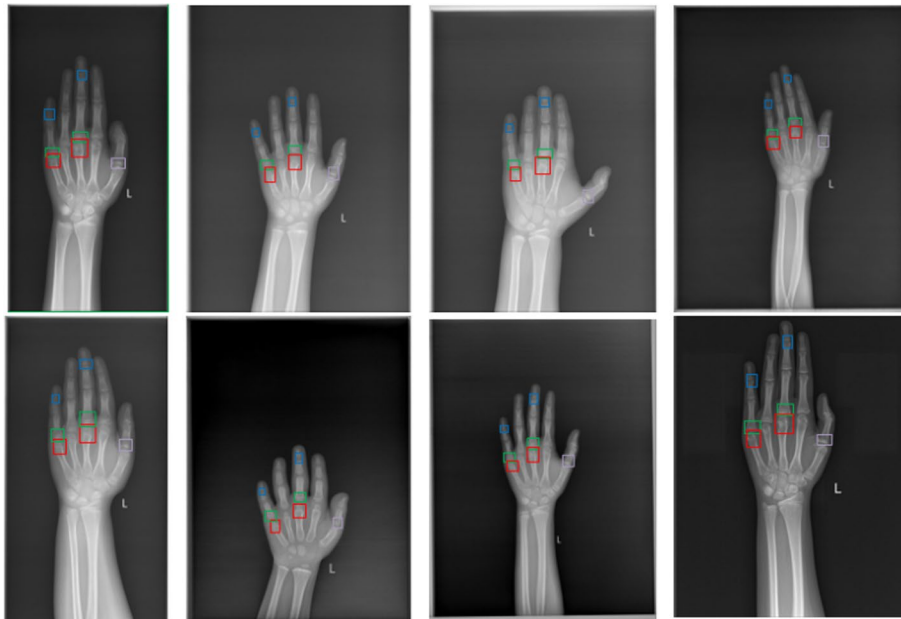


Fig. 11 Samples of annotations of the 3rd metacarpal and 5th metacarpal (red), 3rd proximal phalanx, 5th proximal phalanx (green), 3rd distal phalanx and 5th distal phalanx (blue), and the 1st proximal phalanx (purple). There is an overlap of the red and green boxes, which indicates that the prediction of the 3rd and 5th metacarpal and the 3rd and 5th proximal phalanx will be more difficult than the prediction of the 1st proximal phalanx

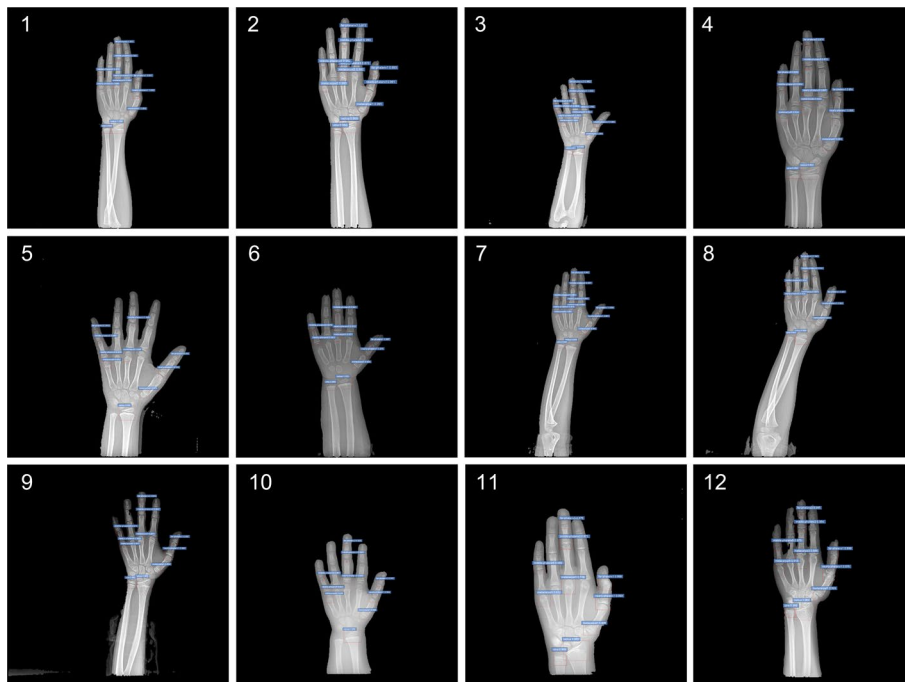


Fig. 12 Example outputs of ROIs detection in hand bone with faster RCNN. The labels of the bounding box, ROI name and AP are shown in the figures

boxes are the 3rd and the 5th metacarpal. We can see that in some cases, the boxes are overlapped, not similar to the 1st proximal phalanx labeled purple, which can be an isolated region. This may be the reason why the AP values are lower for regions 7 and 8. The 3rd distal phalanx and 5th distal phalanx are labeled 12 and 13 in Fig. 2a, they are the blue boxes in Fig. 11. Compared to other ROIs, these two ROIs are smaller, and the detection is harder than other ROIs. Therefore, the APs are lower in other regions.

Example outputs of 13 ROIs detected in hand bone with faster RCNN are shown in Fig. 12. In Fig. 12 (1, 5, 6, 8, 11, 12), 3 ROIs are not detected for each of the images. And no obvious bias can be found in these images, the failure of detection is random. Such APs could not help to mark all the 13 positions accurately. To detect the ROIs of different scales, we adopted YOLO to do multi-scale detection.

Figure 13 shows the example outputs of ROIs detection in hand bone with YOLO. The mAP is 0.9, which is better than the model based on Faster R-CNN. However, except Fig. 13 (5, 7, 12), there are still fail-detected regions in the other 9 images. Therefore, we adopted RFP and SAC in our model to solve this fine-grained multi-objective detection problem. As shown in Fig. 8, all the 13 ROIs can be detected correctly and the locations of the ROIs are also correct, no matter how old the patient is.

4.3 Self-attention transfer network for BAA

In order to avoid over-fitting, we used rotating, scaling, gray-scale transformation to augment the dataset. The loss function used here is cross-entropy. One-hot encoding is used to convert bone age into a format that can be readily used by our algorithm.

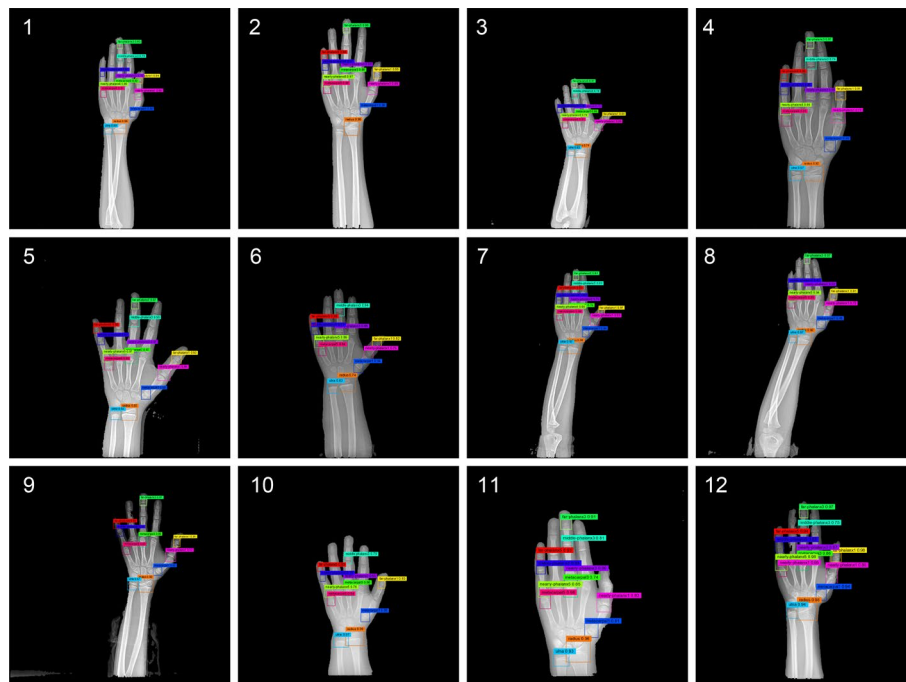


Fig. 13 Example outputs of ROIs detection in hand bone with YOLO. The labels of the bounding box, ROI name and AP are shown in the figures

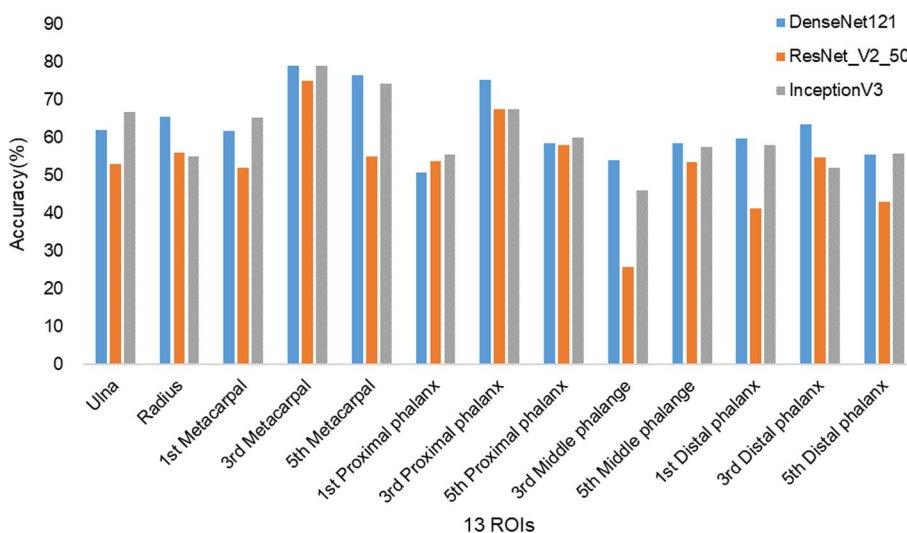


Fig. 14 Test result of the BAA of all ROIs

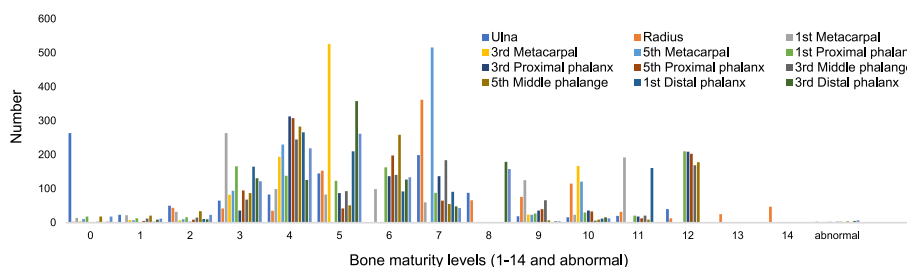


Fig. 15 Statistics of samples at different levels (totally 873 samples in the dataset)

Because the numbers of the samples at different levels vary greatly, class weight is used during the model training. We set learning rate 0.01, 0.001 and 0.005. The batch size is set 2, the momentum is 0.9. SGD is used as the optimizer. When the epoch is 100, the learning rate is 0.01, and we get the best performance of the model. The accuracy of the bone maturity assessment is shown in Fig. 14.

DenseNet121, ResNet_V2_50 and Inception-V3 were used for the maturity level prediction of the ROIs and the mean accuracy is 63.15%, 52.97% and 61.74%. DenseNet121 shows the best performance and it is used in our BAA framework. We can see that the accuracy for the 1st proximal phalanx and the 3rd middle phalange are lower than other ROIs, because in TW method, the level of the maturity is on a finer scale, which leverages the resolution of the bone maturity. However, on the other side, the difference between levels will be smaller and the accuracy of the assessment will be lower. There are 15 levels for radius; 13 levels for ulna, 1st proximal phalanx, 3rd proximal phalanx, 5th proximal phalanx, 3rd middle phalange and 5th middle phalange; 12 levels for 1st metacarpal, 1st distal phalanx, 3rd distal phalanx and 5th distal phalanx; 11 levels for 3rd metacarpal and 5th metacarpal. For one ROI, getting enough data for training is quite challenging. The statistics of the ROI samples are in Fig. 15. For example, there are only 3 samples for the radius at level 0, no samples at

Table 2 Test result of the BAA

Age group	MAE (year)	RMSE (year)
0–3	0.95	0.99
4–7	0.72	0.78
8–11	0.48	0.54
12–15	0.57	0.62
16–18	0.94	1.22
ALL (0–18)	0.61	0.69
ALL (2–18)	0.59	0.67

level 6; for the 3rd metacarpal, no samples at levels 6, 7, and 8. It is believed that the imbalance and the limited scale of the dataset is one of the reasons for the accuracy of 63.15%.

Mean average error (MAE) is used to evaluate the final estimated bone age. It is calculated as Eq. (4) shows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |m_i - e_i|. \quad (4)$$

m_i is the BAA with our algorithm, and e_i is the ground truth labeled by the professional radiologists. 222 images have been tested with the model and the results are shown in Table 2.

It is shown that in the age groups 0–3, 4–7 and 16–18, the MAEs are larger than other age groups, because the cases under 7 years and older than 16 years are less than other groups. Finally, we obtained a MAE of 0.61 year (7.32 months) on our dataset. The resolution of the label of the bone age in our dataset is 1 year, while for some public datasets, the resolution is months. It is hard to compare the different models fairly. However, we presented some of the published results in Table 3 to show the differences between these studies.

Table 3 is a summary of the bone age testing methods based on deep learning. In [3, 10, 28–30], the MAEs of the end-to-end models are better than the subregion-based method in [9, 20] and this paper. End-to-end BAA is a multi-class regression task, and it has lower computation complexity and better MAE comparing to the subregion-based method. However, just giving a bone age is hard to satisfy the clinical requirement now. Son et al. [32] detect 13 ROIs using Faster RCNN and the rejection rate of the approach is approximately 1.6%, which means the rejected X-ray image cannot be evaluated by the proposed model. They latter used a correlation matrix to convert the 13 maturity predictions into a single bone age. They reported an MAE of 5.52 months on their internal test set, while the rejected X-ray images were not counted in this evaluation. Bui et al. [20] detect 6 ROIs using a Faster R-CNN detection network, and they train a support vector regression model for BAA. Their MAE is similar to our result. However, they just provide diagnostic information of 6 regions, which is not consistent with the TW method (13 ROIs) and less information was given. In contrast to their works, our work is capable to detect the 13 ROIs steadily and it is not needed to worry about the limitation that the assessment will fail if not all ROIs are detected [21]. This is a big progress in region-based BAA, and the performance is higher than all the state-of-the-art BAA methods.

Table 3 Summary of the bone age testing methods based on deep learning

Category of the method	References	Method	Dataset	Resolution of label	MAE (month)
End-to-end	Alexander et al. [28]	Inception-V3 + DenseNet; fine-tuned ResNet-50; ice module; U-Net segmentation and CNN-based recognition	RSNA dataset Total data: 14,236 Training set: validation set: test set = 12,611:1425:200 The ratio of male to female is close to 1:1	1 month	4.27–4.5 months
	Ren et al. [3]	Regression CNN based on inception-V3	RSNA dataset SCH data set (Shanghai Children’s Hospital): total data: 12,390 Training set: validation set: test set = 9912:1239:1239	1 month	5.2 months
	Chen et al. [10]	ResNet + spatial transformer + LBP + GLCM + SVM	Total data: 12,536 from Shengjing Hospital, Training set: test set = 85:15	Not mentioned	5.5 months
	Mutasa et al. [29]	ResNet and inception	Total data: 20,581 (10,289 from the network, 8909 from Columbia University Hospital, 1383 from the public data set) Training set: validation set: test set = 11,007:1105:300	1 month	6.43 months
	Liu et al. [30]	NSCT-based multi-scale CNN (VGG16)	Digital Hand Atlas Database System Total data: 1391	12 months (1 year)	6.43 months
	Iglovikov et al. [47]	U-Net + VGG-style + linear combination	RSNA dataset	1 month	7.52 months
	Hu et al. [48]	AlexNet	DR images of Uyghur people, total data: 472 Training set: test set = 7:3	1 month	8.40 months
	Lee et al. [16]	Caffenet architecture	Total data: 600 Training set: test set = 2:1	Not mentioned	18.9 months
	Synho et al. [18]	CNN segmentation + Lenet BAA	Training set: verification set: test set = 70:15:15	12 months (1 year)	Accuracy of MAE < 1 year: 92.29%
	Tajmir et al. [17]	LeNet-5	Total data: 8325 Training set: test set = 8045:280	12 months (1 year)	Accuracy of MAE < 1 year: 98.6%
Subregion-based	Son et al. [32]	faster R-CNN + VGG	Total data: 3344 Training set: test set = 4:1 Age: 2–14 years old	Not mentioned	5.52 months Reject rate: 1.6%
	Bui et al. [20]	faster R-CNN + TW3	Total data: 1375 Training set: test set = 4:1	12 months (1 year)	7.08 months
	Spampinato et al. [9]	BoNet (ad-hoc CNN, 6 layers)	Digital Hand Atlas Database System Total data: 1391	12 months (1 year)	9.48 months
	Proposed method	U-Net + RPN + cascade R-CNN + DenseNet with self-attention	Total data: 2129 Annotated for fine-grained BAA: 873 Training set: test set = 4:1 Age: 2–18 years	12 months (1 year)	7.32 months (7.08 months for age 2–18) Reject rate: 0

5 Discussion

Following most machine learning-based BAA studies, most of them are end-to-end deep learning based. In practice, radiologists and doctors want more detailed information about the variation of the bones in hand, rather than age. More recently, several sub-region-based BAA methods were proposed. The ROIs in hand are extracted and evaluated firstly and then with a regression model the bone age can be obtained. However, if the ROI detection failed, the latter processing will be difficult. And Fast R-CNN is used in most of these works, and missing detection exists in all these studies [9, 20, 32]. In addition, to promote the detection performance, the TW method was simplified and only 6 ROIs were considered in [20] and not all the diagnosis information can be provided for doctors. A robust and precise BAA method is desired. For this purpose, a four-stage fine-grained BAA method that includes palm segmentation, 13-ROI detection, fine-grained ROI maturity assessment, and precise-BAA is proposed. The four parts are independent and concurrent, which means we can develop each part separately to optimize the performance of BAA. This design could potentially improve the BAA study further. Researchers could focus on and contribute to one part they are interested in, such as segmentation or comprehensive assessment.

In the image preprocessing stage, a palm segmentation method based on U-Net, hole filling, and connected component selection is proposed. Compared with the threshold segmentation and level set [24], the accuracy of the segmentation is better and can be improved to 98.01%. The consistency of the images brought a low rejection rate, which laid a solid foundation for ROI detection. In the stage of multi-target detection, a recursive feature pyramid-based model is built to handle the fine-grained ROIs detection. Compared to the most commonly used models including Fast R-CNN and YOLO, we get the state-of-the-art result in multi-bone detection. In the stage of ROI based skeletal maturity evaluation based on transfer learning, a dataset with 1015 images is built. As far as we know, this is the largest finely annotated dataset for BAA, which includes 1015×13 ROIs. With a prediction model fine-tuned with our own dataset, we can get the skeletal maturity level of each ROI in each X-ray image. In the last stage of bone age calculation, the bone age is calculated based on the 50th percentile curve of bone maturity, and we can get an MAE of 7.32 months for age 1–18 (7.08 months for age 2–18). However, it is hard to compare the effectiveness of the proposed procedures. We found that the scale of the dataset, age distribution, valid range of age in MAE calculation would affect the result tremendously. Nevertheless, with more data, the performance of the model will be better.

At present, the software of BAA based on the whole atlas has been used in some institutions. However, for fine-grained skeletal maturity assessment, the performance of the models still needs to be further improved. In the third stage of this study, the maturity assessment of the ROIs in hand bone, the resolution of our dataset is 1 year, which is larger than many other datasets [3, 28, 29, 47, 48]. The low resolution may cause a larger MAE. A better and larger dataset should be built for subregion-based research. And we will collaborate with doctors build a larger and high-resolution dataset for a precise-BAA study.

6 Conclusion

Nowadays, BAA has already become a common clinical examination. However, at present, in major domestic medical institutions, BAA is still based on artificial evaluation, and it takes a lot of time for doctors or radiologists to assess bone age subjectively. In the past 2 years, most of the studies in the automatic BAA are using a whole hand bone X-ray image in an end-to-end way. While it only gives a number as the bone age and it is impossible to see the development of each bone in hand, and the model lacks interpretability. It is meaningful and necessary to study the fine-grained BAA method based on deep learning. Radiologists can check the development of the main hand bone areas, and give diagnosis and treatment on time. With our four-stage framework, we can get the segmented palm, the 13 ROIs in the hand area, skeletal maturity of each ROI, and the final bone age of the case. This fine-grained BAA model provides more details about the development of the case for doctors and radiologists. We believe it is more valuable to analyze the patient's condition development and adjust the treatment plan in time.

Abbreviations

BAA	Bone age assessment
ROIs	Region of interests
MAE	Mean absolute error
CAD	Computer-aided diagnosis
TW method	Tanner–Whitehouse method
RUS-CHN	RUS-CHN radiographic atlas method
G-P method	Greulich–Pyle method
CHN	The Skeletal Development Standards of Hand and Wrist for Chinese Children—China 05
ASM	Active Shape Model
RSNA	The North American Radiology Society
RPN	Region Proposal Network
DIP	Distal interphalangeal joint
PIP	Proximal interphalangeal joint
MCP	Metacarpophalangeal joint
IoU	Intersection over union
RFP	Recursive feature pyramid network
ASPP	Atrous spatial pyramid pooling
SAC	Switchable atrous convolution
SGD	Stochastic gradient descent
AP	Average precision
mAP	Mean average precision
GIoU	Generalized intersection over union
YOLO	You only look once, real-time object detection
COCO	Common objects in context dataset
VOC	Visual object classes dataset
RMSE	Root mean square error

Acknowledgements

We acknowledge Jungang Han, Yufeng Lu for useful discussions, and Xiaochen Niu for data storage service.

Author contributions

YJ performed the data analyses and wrote the manuscript; WC performed the experiment; XZ, HD contributed significantly to analysis and manuscript preparation; XJ, WQ contributed to the conception of the study; BY, QZ contributed to the dataset preparation; ZW helped perform the analysis with constructive discussions. All authors read and approved the final manuscript.

Funding

This research is supported by Key Research and Development Program of Shaanxi Province (2019GY-021), Science and technology program of Xi'an, (GX YD17.12), Open fund of Shaanxi Key Laboratory of Network Data Intelligent Processing (XUPT-KLND (201802, 201803)).

Data availability

The X-ray images of hand bone that support the findings of this study are available to download: http://222.24.63.112:8160/jiayang/hand_bone_X-ray.zip

Code availability

The source code, weights of trained models used for experiments of precise-BAA are available in GitHub: https://github.com/dhr079/bone_age_automatic_assessment.

Declarations**Competing interests**

The authors declare no competing interests.

Received: 13 September 2021 Accepted: 4 July 2022

Published online: 26 July 2022

References

1. D.D. Martin, J.M. Wit, Z. Hochberg, L. Säwendahl, R.R. van Rijn, O. Fricke et al., The use of bone age in clinical practice—part 1. *Horm. Res. Paediatr.* **76**, 1–9 (2011)
2. Y. Yan, Research on ID identification and bone age testing of 2011 Double Happiness-New Star Cup National Children's table tennis players, Master, Soochow University, 2012
3. X. Ren, T. Li, X. Yang, S. Wang, S. Ahmad, L. Xiang et al., Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph. *IEEE J. Biomed. Health Inform.* **23**(5), 2030–2038 (2018)
4. T.W. Todd, Atlas of skeletal maturation. *J. Anat.* **72**, 640 (1938)
5. S. Zhang, *The skeletal development standards of hand and wrist for Chinese children—China 05 and its application* (China Science Publishing & Media Ltd, Beijing, 2015)
6. H. Goldstein, N. Cameron, J.M. Healy, M. Tanner, Assessment of skeletal maturity and predication of adult height (TW3 method). *Gov. Oppos.* **36**, 27–47 (2001)
7. BoneXpert. <https://www.bonexpert.com/>
8. H.H. Thodberg, S. Kreiborg, A. Juul, K.D. Pedersen, The BoneXpert method for automated determination of skeletal maturity. *IEEE Trans. Med. Imaging* **28**, 52–66 (2009)
9. C. Spampinato, S. Palazzo, D. Giordano, M. Aldinucci, R. Leonardi, Deep learning for automated skeletal bone age assessment in X-ray images. *Med. Image Anal.* **36**, 41–51 (2017)
10. X. Chen, J. Li, Y. Zhang, Y. Lu, S. Liu, Automatic feature extraction in X-ray image based on deep learning approach for determination of bone age. *Future Gen. Comput. Syst.* **110**, 795–801 (2020)
11. Y. Wang, Q. Zhang, J. Han, Y. Jia, Application of deep learning in bone age assessment. Presented at the IOP conference series: earth and environmental science (EES), Guangzhou, 2018
12. J. Han, Y. Jia, C. Zhao, F. Gou, Automatic bone age assessment combined with transfer learning and support vector regression, in *2018 9th international conference on information technology in medicine and education (ITME)*, 2018
13. B.Y. Yonggang Tang et al., Assessment and analysis of wrist bone age in 190 adolescents in Xi'an. *Shaanxi Med. J.* **47**, 1661–1663 (2018)
14. C. Zhao, J. Han, Y. Jia, L. Fan, F. Gou, Versatile framework for medical image processing and analysis with application to automatic bone age assessment. *J. Electr. Comput. Eng.* (2018). <https://doi.org/10.1155/2018/2187247>
15. Y. Jia, H. Du, H. Wang, W. Chen, X. Jin, W. Qi, B. Yang, Q. Zhang, A survey of deep learning based fully automatic bone age assessment algorithms. Presented at the pattern recognition. ICPR international workshops and challenges, 2021
16. J.H. Lee, K.G. Kim, Applying deep learning in medical images: the case of bone age estimation. *Healthc. Inform. Res.* **24**, 86 (2018)
17. S.H. Tajmir, H. Lee, R. Shailam, H.I. Gale, J.C. Nguyen, S.J. Westra et al., Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability. *Skelet. Radiol.* **48**, 275–283 (2019)
18. H. Lee, S. Tajmir, J. Lee, M. Zissen, B.A. Yeshiwas, T.K. Alkasab et al., Fully automated deep learning system for bone age assessment. *J. Digit. Imaging* **30**, 427–441 (2017)
19. B. Liang, Y. Zhai, C. Tong, J. Zhao, J. Li, X. He et al., A deep automated skeletal bone age assessment model via region-based convolutional neural network. *Future Gen. Comput. Syst.* **98**, 54–59 (2019)
20. T.D. Bui, J.J. Lee, J. Shin, Incorporated region detection and classification using deep convolutional networks for bone age assessment. *Artif. Intell. Med.* **97**, 1–8 (2019)
21. S. Koitka, M.S. Kim, M. Qu, A. Fischer, F. Nensa, Mimicking the radiologists' workflow: estimating pediatric hand bone age with stacked deep neural networks. *Med. Image Anal.* **64**, 101743 (2020)
22. A. Wibisono, P. Mursanto, Multi region-based feature connected layer (RB-FCL) of deep learning models for bone age assessment. *J. Big Data* **7**, 1–17 (2020)
23. S. Zhang, L. Liu, The skeletal development standards of hand and wrist for Chinese children—China 05 I. TW_3-C RUS, TW_3-C Carpal, and RUS-CHN Methods. *Chin. J. Sports Med.* 6–13 (2006)
24. S. Simu, S. Lal, A study about evolutionary and non-evolutionary segmentation techniques on hand radiographs for bone age assessment. *Biomed. Signal Process. Control* **33**, 220–235 (2017)
25. C.H. Yan, *Segmentation of Hand Bone for Bone Age Assessment* (Springer, Cham, 2013)
26. C. Yan, B. Gong, Y. Wei, Y. Gao, Deep multi-view enhancement hashing for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 1445–1451 (2020)
27. C. Yan, Z. Li, Y. Zhang, Y. Liu, X. Ji, Y. Zhang, Depth image denoising using nuclear norm and learning graph model. *ACM Trans. Multimed. Comput. Commun. Appl.* **16**, 1–17 (2020)
28. S.S. Halabji, L.M. Prevedello, J. Kalpathy-Cramer, A.B. Mamonov, A. Bilbily, M. Cicero et al., The RSNA pediatric bone age machine learning challenge. *Radiology* **290**, 498–503 (2018)

29. S. Mutasa, P.D. Chang, C. Ruzal-Shapiro, R. Ayyala, MABAL: a novel deep-learning architecture for machine-assisted bone age labeling. *J. Digit. Imaging* **31**, 513–519 (2018)
30. Y. Liu, C. Zhang, J. Cheng, X. Chen, Z.J. Wang, A multi-scale data fusion framework for bone age assessment with convolutional neural networks. *Comput. Biol. Med.* **108**, 161–173 (2019)
31. M. Jaderberg, K. Simonyan, A. Zisserman, Spatial transformer networks, in *Advances in Neural Information Processing Systems*. (MIT Press, Cambridge, 2015), pp. 2017–2025
32. S.J. Son, Y. Song, N. Kim, Y. Do, N. Kwak, M.S. Lee et al., TW3-based fully automated bone age assessment system using deep neural networks. *IEEE Access* **7**, 33346–33358 (2019)
33. S. Koitka, A. Demircioglu, M.S. Kim, C.M. Friedrich, F. Nensa, Ossification area localization in pediatric hand radiographs using deep neural networks for object detection. *PLoS ONE* **13**, e0207496 (2018)
34. Y. Jia, W. Chen, M. Yang, L. Wang, D. Liu, Q. Zhang, Video smoke detection with domain knowledge and transfer learning from deep convolutional neural networks. *Optik* **240**, 166947 (2021)
35. O. Ronneberger, P. Fischer, and T. Brox, U-net: convolutional networks for biomedical image segmentation, in *International conference on medical image computing and computer-assisted intervention*, 2015, pp. 234–241
36. S. Qiao, L. C. Chen, A. Yuille, DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution (2020), Preprint at [arXiv:2006.02334](https://arxiv.org/abs/2006.02334)
37. L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2018)
38. T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, IEEE, 2016
39. Z. Cai, N. Vasconcelos, Cascade R-CNN: high quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(5), 1483–1498 (2019)
40. J. Yosinski, J. Clune, Y. Bengio, H. Lipson, *How Transferable are Features in Deep Neural Networks?* (MIT Press, Cambridge, 2014)
41. C. Yan, Y. Hao, L. Li, J. Yin, A. Liu, Z. Mao et al., Task-adaptive attention for image captioning. *IEEE Trans. Circuits Syst. Video Technol.* **32**, 43–51 (2021)
42. C. Yan, L. Meng, L. Li, J. Zhang, Z. Wang, J. Yin et al., Age-invariant face recognition by multi-feature fusion and decomposition with self-attention. *ACM Trans. Multimed. Comput. Commun. Appl.* **18**, 1–18 (2022)
43. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez et al., Attention is all you need, in *NIPS*, 2017
44. H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (2019), pp. 658–666
45. R. Girshick, Fast R-CNN, in *Proceedings of the IEEE international conference on computer vision*, 2015
46. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149 (2017)
47. V.I. Iglovikov, A. Rakhlin, A.A. Kalinin, A.A. Shvets, Paediatric Bone age assessment using deep convolutional neural networks, in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. (Springer, Cham, 2018), pp. 300–308
48. T.H. Hu, Z. Huo, T.A. Liu et al., Automated assessment for bone age of left wrist joint in Uyghur teenagers by deep learning. *J. Forensic Med.* **34**, 27–32 (2018)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
